Concentration of Measure for Computational Science

LDRD ER-MSM

Name	Clint Scovel			**Z Number**	097403	**Group**	CIC-3
Phone	5-4721	**FAX**	5-5220	**Mail Stop**	B265	**E- mail**	jcs@lanl.gov

1 Executive Summary

The probabilistic method is designed to replace *worst-case* with *typical* in the analysis of computational problems. In the past decade it has revolutionized computational science largely because of the inability of conventional worst-case analysis to adequately characterize important computational problems. Recent advances in the theory of concentration of measure, such as the works of Talagrand, McDiarmid, Marton, and Boucheron-Lugosi-Massart, have played a central role. The theory of concentration of measure is the science of characterizing subsets where the probability mass is concentrated and therefore represents typical behavior.

Computational science at LANL is also undergoing a revolution due to the computational demands required to solve scientific problems of national importance. Clearly the theory of concentration of measure will play a central role here also. *We propose to develop the theory of concentration of measure to facilitate the revolution of computational science at LANL.*

2 Background

Consider the problem of sorting a sequence of n integers in increasing order utilizing the comparison operation between pairs of numbers. Algorithms exist whose worst-case run time achieves the theoretical lower bound of $n \log n$ for this problem. However, even though its worst-case run time is n^2, Hoare's Quicksort algorithm is often preferred because of its simplicity and the fact that it frequently yields better run times. This suggests that a worst-case analysis of Quicksort is not indicative of its true performance. Knuth was the first to resolve this issue by showing that if one places a reasonable probability distribution on the space of sequences, then the run time of Quicksort is $n \log n$ with high probability. Knuth's result not only provides a theoretical justification for Quicksort's superior performance, but also demonstrates the utility of the probabilistic method pioneered by Erdös. The probabilistic method has not only provided a better understanding of numerous important computational problems, it has revolutionized computational science. In particular it has provided new tools for the characterization of computational problems, explanations for the performance of existing algorithms, and a framework for the development of new algorithms.

The probabilistic method is designed to replace *worst-case* with *typical* in the analysis of computational problems. This is accomplished by the introduction of a random

variable and the application of probability theory to obtain results formulated as probabilistic statements. One common method introduces a probability distribution on the space of problem instances. This method can be used to both characterize the intrinsic properties of a computational problem (e.g. bounding the chromatic number of a graph), and to analyze the computational complexity of solution algorithms (e.g. characterizing the run time of the Kernighan-Lin algorithm for TRAVELING SALESMAN). Another method, randomized algorithms, introduces a random process in the algorithm itself.

Having a probability distribution on some primal aspect of the computational problem induces a probability distribution on the *objective variable* through its *functional relationship* to the *primal variable*. For example the distribution of the objective variable *run time* for HAMILTONIAN CYCLE is the result of propagating the distribution of the primal variable *edge configurations* through the functional relationship determined by an *algorithm*. The most fundamental statistic of the objective variable is its expected value. Depending on the nature of the relationship between the objective variable and the primal random variable the theory of *concentration of measure* can often show that the distribution of the objective variable is concentrated about its expected value. In this case the expected value represents the typical value of the objective variable. In addition, this result is often robust to the choice of primal distribution so that the expected value provides a bound for the value of the objective variable observed in practice.

The theory of concentration of measure, described below, is the tool that has enabled the probabilistic method to successfully attack many outstanding computational problems. For example Frieze and Reed [4] have used it to constructed a polynomial-time algorithm that with high probability solves the NP-complete problem HAMILTONIAN CYCLE on random graphs with edge probability 1/2. Thus, even though the problem has hard instances, the typical instance is easy. They also use concentration of measure to show that a certain branch and bound algorithm for KNAPSACK takes super-polynomial time with high probability. In this case the typical instance is hard, even though there are easy instances. Consequently the utilization of concentration of measure in the probabilistic method has revolutionized the study of these landmark problems by providing resolutions to issues which have eluded researchers for several decades. More generally the theory of concentration of measure has become an invaluable tool in discrete mathematics, combinatorics, probabilistic analysis of algorithms, analysis of randomized algorithms, and machine learning, and therefore has evolved into a major research area in probability. For recent surveys see Alon, et.al. [1] and Habib, et.al. [5].

The theory of concentration of measure is the science of characterizing subsets where the probability mass is concentrated. The most common example is proving when a random variable is concentrated about its expected value. Note that the terminology "expected value" is misleading in that it suggests that this value is typical. For example Figure 1 shows the distribution of run times for some algorithm. Clearly the expected value E is not typical of the run time and seriously misrepresents the performance of this algorithm. On the other hand Figure 2 shows the distribution of run times for an algorithm which possesses a concentration result. In this case the probability distribution is concentrated about its expected value and therefore the expected value faithfully

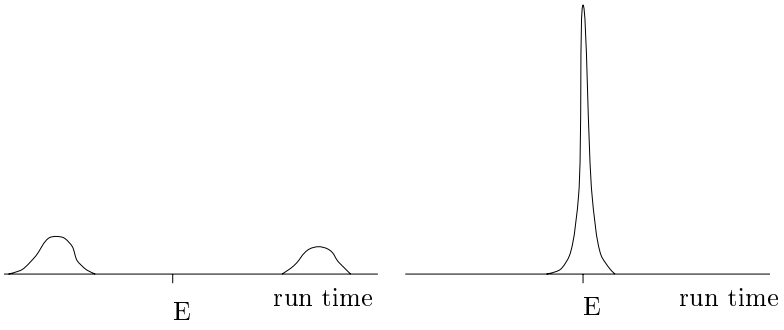

Figure 1

Figure 2

represents the typical run time.

In applications the functional relationship between the objective and primal variables may be unknown or complicated to compute. For example consider the problem of computing the chromatic number of a random graph where the primal variable is the set of edges and the objective variable is the chromatic number. In such cases even when the distribution of the primal random variable is known, the distribution of the objective variable is not and so the concentration of the objective variable about its expected value may not be addressed directly. The theory of concentration of measure addresses this situation by developing results using known properties of both the functional relationship and the distribution of the primal random variable. Recent work has shown that positive results can be obtained for primal distributions and functional relationships for which only generic properties are known. For example Talagrand's concentration inequality has transformed this field and has pioneered the way for its future development [10].

The weakest concentration result is Chebychev's inequality,

$$P(|X - E[X]| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

where σ^2 is the variance of the random variable X. For small σ, X is concentrated around $E[X]$, but the bound is weak because it is polynomial in $1/\epsilon$. The first strong concentration result is Chernoff's inequality,

$$P(|\bar{X} - E[X]| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of n independent identically distributed (iid) binomial random variables. When $E[X] = 1/2$ Chebychev's inequality concentrates $\bar{X}$ by $\delta = \frac{1}{4n\epsilon^2}$ while Chernoff's inequality concentrates it by the exponentially stronger $2e^{-1/2\delta}$. Note that these bounds are not asymptotic and hold strictly for any finite n. The non-asymptotic nature of these bounds is maintained in all the following concentration results. Indeed this non-asymptotic property is a distinguishing feature of these bounds and is essential to the practitioner working on finite problems. Hoeffding extended Chernoff's inequality to general iid random variables. Recent theoretical advances provide concentration results for more general classes of functions than the sample mean. At present

there appears to be four main techniques, all making different assumptions on the class of functions; McDiarmid's inequality based on martingale differences [10], Talagrand's inequality based on his induction method [10], Marton's inequality based on information theory [9], and the inequality of Boucheron, Lugosi and Massart [2] utilizing logarithmic Sobolev inequalities. These techniques collectively cover a broad class of functions and therefore are applicable to many computing problems. On the other hand these results have been formulated without concern for specific computational problem domains, and computer scientists have simply utilized the results that most closely match their needs. Consequently for many computing problems concentration results do not yet exist, while for others existing results are not as powerful as they could be. *We propose to address this shortcoming by tailoring concentration theory to specific problem domains.*

3 Proposed Work and Importance to LANL

In the past decade computational science has been revolutionized by the probabilistic method largely because of the inability of conventional worst-case analysis to adequately characterize important computational problems. Recent advances in the theory of concentration of measure have played a central role. Computational science at LANL is also undergoing a revolution due to the computational demands required to solve scientific problems of national importance. Clearly the theory of concentration of measure will play a central role here also. We propose to develop the theory of concentration of measure to facilitate the revolution of computational science at LANL. To accomplish this task we proceed on two fronts; the extension of the general theory and the development of a theory for specific problem domains. We have made progress on both fronts in machine learning. In particular in Hush and Scovel [7] we extended the general result of Boucheron, et.al., while in Hush and Scovel [6] we improve the specific result of Koltchinskii [8] concerning the concentration of the Rademacher statistic. In summary

- we propose to extend the general theory of concentration of measure represented by the works of Talagrand, McDiarmid, Marton, and Boucheron-Lugosi-Massart.
- we propose to develop a theory of concentration of measure for specific computational problem domains at LANL.

Many of the scientific problems studied at LANL contain uncertainties that can be modeled using primal random variables which are propagated through complex systems to yield the objective random variable. In many cases it is possible to obtain reasonable estimates of the expected value of the objective variable. If we wish to interpret this expected value as a faithful representative of the object of study, concentration of measure is required. Examples of LANL programs which can benefit from this research include parameter estimation in weapons design codes, predictability, monitoring the health of nuclear weapons, detection of hard and deeply buried targets, simulations of large scale socio-technical systems, network intrusion detection, automatic document classification, and genome sequencing. We describe a version of predictability in some detail. Here

the goal is to determine the accuracy of numerical simulations in approximating the evolution of a field equation. The technique of optimal prediction invented by Chorin, et.al. [3] places a probability distribution on the space of fields at time zero conditioned on the value of the discretized field at time zero. Simulations are performed which are designed to approximate the evolution of the field equation. They are designed to accurately propagate the expected value, but their propagation of the distribution is not clear. To interpret the expected value as an accurate representation of the solution requires the concentration of the propagated probability measure about this value. To put this in the context of this proposal, let the field at time zero be the primal variable distributed according to the conditional probability distribution at time zero. Let the objective variable be the field at time t distributed according to the propagated pdf. Then the functional relationship between the primal and objective variable used in the study of concentration of measure is determined by the simulation. By considering generic properties of the simulation it may be possible to establish concentration of measure, thereby providing confidence in the simulation's predictions. Success in these types of efforts would provide confidence in the accuracy of simulation methods used at LANL on problems of national importance.

4 Specialist Reviewers

Ken Hanson, DX-3
Madhav Marathe, TSAS
Vladimir Koltchinskii, Department of Mathematics, University of New Mexico, vlad@math.unm.edu
Colin McDiarmid, Department of Statistics, University of Oxford, cmcd@stats.ox.ac.uk

5 Funding Breakout and Key Participants

Key Technical Staff: Clint Scovel (CIC-3), Don Hush (CIC-3).

Funding: \$150K of funding is requested for each of 3 years. All funding is for labor, Scovel and Hush at approximately 0.5 FTE each.

References

[1] Alon, N., Spencer, J. H., and Erdös, P., *The Probabilistic Method*, John Wiley, New York, 1992.

[2] Boucheron, S., Lugosi, G., and Massart, P., A sharp inequality with applications, preprint, 1999.

[3] Chorin, A.J., Kast, A.P., and Kupferman, R., Optimal prediction for underresolved dynamics, *Proc. Nat. Acad. Sci. USA*, **95**, pp. 4094-4098, 1998.

[4] Frieze, A. M. and Reed, B., Probabilistic Analysis of Algorithms, *Probabilistic Methods for Algorithmic Discrete Mathematics* Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., Reed, B., Eds., pp. 36–92, Springer-Verlag, Berlin, 1998.

[5] *Probabilistic Methods for Algorithmic Discrete Mathematics*, Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., Reed, B., Eds., Springer-Verlag, Berlin, 1998.

[6] Hush, D., Scovel. C., Conditional performance bounds for machine learning, submitted to *Machine Learning*, 1999.

[7] Hush, D., Scovel. C., A new proof of concentration of Rademacher statistics, submitted to *Annals of Probability*, 1999.

[8] Koltchinskii, V. I., Rademacher Penalties and Structural Risk Minimization, preprint, 1999.

[9] Marton, K., Bounding $\bar{d}$-Distance by informational divergence: A method to prove measure concentration, *The Annals of Probability* **24**(1996), 857–866.

[10] McDiarmid, C.,, Concentration, *Probabilistic Methods for Algorithmic Discrete Mathematics* Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., Reed, B., Eds., pp. 195–248, Springer-Verlag, Berlin, 1998

James C. Scovel
Los Alamos National Laboratory
Computer Research Group, CIC-3
MS B265, Los Alamos, NM 87545
(505) 665-4721, jcs@lanl.gov, http://cnls.lanl.gov/ jcs

FORMAL EDUCATION:

Ph.D.	Mathematics, Courant Institute of Mathematical Sciences, 1983
M.S.	Mathematics, University of Arizona, 1979
B.S.	Mechanical Engineering, Cornell University, 1977

RESEARCH INTERESTS:

Machine Learning and Concentration of Measure.

EMPLOYMENT HISTORY:

1989–Present	Staff Scientist, Computer Research Group, CIC-3
1986–1989	Staff Scientist, Mathematical Modeling and Analysis Group, T-7
1983–1986	Assistant Professor of Mathematics, Brandeis University
1989	Visiting Scientist, Mathematical Sciences Institute, Cornell University
1989	Academic Guest, Forschungsinstitut für Mathematik, ETH-Zürich, Switzerland

PROFESSIONAL ACTIVITIES (1996-99):

1996–1998	Technical Lead for the HCFA Medicare Fruad Detection Project at LANL

PATENTS:

1. (1998) with J. Hogden, and J. White, *Anomaly Analysis Using Maximum Likelihood Continuity Mapping* S-87,239, U.S. Patent 6,038,388.

SELECTED PUBLICATIONS (1996-99):

1. "An equivalence relation between parallel calibration and principal component regression, with R. Christensen, M. Fugate, and D. Hush, submitted to *Journal of Chemometrics*, 2000.

2. "Conditional performance bounds for machine learning," with Don Hush, submitted to *Machine Learning*, 1999.

3. "A new proof of concentration of Rademacher statistics," with Don Hush, submitted to *Annals of Probability*, 1999.

4. "On the VC Dimension of Bounded Margin Classifiers," with Don Hush, to appear in Machine Learning, 2000, LA-UR-99-2526.

5. "Logistic Regression with Incomplete Choice-Based Samples," with M. Fugate, and A. Marathe, submitted to Communications in Statistics - Theory Meth. 1998.

6. "Bayesian Stratified Sampling to Assess Corpus Utility," with Hochberg, J., Thomas, T., and Hall, S., Proceedings of the Sixth Workshop on Very Large Corpora, E. Charniak, Ed., San Francisco, CA, Morgan Kaufmann, 1998, pp. 1–8, LA-UR-98-1922

7. "Disaggregating Time Series Data," with T. Burr, and S. B. Joubert, 1997, LANL report LA-13292-MS.

8. "Fraud Detection in Medicare Claims: A Multivariate Outlier Detection Approach," with T. Burr, C. Hale, M. Kantor, D. Weiss, J. White, 1997, LA-UR-97-1142.

9. "Comparing Candidate Hospital Report Cards," Proceedings of the American Statistical Society Joint Statistical Meetings, Statistical Graphics Section, Anaheim Ca, Aug 11-14, 1997, p 112-115, LA-13293-MS.

10. "Improving Prediction by Linear Combination of Generalizers with Given Smoothness ," with G. P. Berman, and G. V. Lopez, 1996, LAUR=?,

11. "Los Alamos National Laboratory's Contribution to the NQR Test," with Hochberg, J., P. Fasel, & T. Tillmann, 1996. LA-CP #97-67.

12. "Knowledge fusion: An approach to time series model selection followed by pattern recognition," with S. Bleasedale, Strittmatter, R., and T. L. Burr, 1996, LAUR report LA-13095-MS.

13. "Graph-Theoretical Approaches to Detecting Ping-Pong Schemes," with A. Katsevich, 1996, unpublished report submitted to HCFA medicare organization.

14. "Chartrand's Theorem for Bipartite Graphs," 1996, LA-UR-96-2721.

Don R. Hush

Address

Group CIC-3	Phone: 505-665-2722
MS B265	FAX: 505-665-5220
Los Alamos National Laboratory	Email: dhush@lanl.gov
Los Alamos, NM 87545	

Education

PhD Elec Eng	1986	University of New Mexico
MS Elec Eng	1982	Kansas State University
BS Elec Eng	1980	Kansas State University (Summa Cum Laude)

Professional Experience

1998–present	Technical Staff Member, Los Alamos National Laboratory
1993–1998	Associate Professor, University of New Mexico
1987–1993	Assistant Professor, University of New Mexico
Summer 1991,1994,1996	Visiting Professor, Universidad de Vigo, Vigo, Spain
1986–1987	Technical Staff Member, Sandia National Laboratories
1994–1997	Associate Editor, IEEE Transactions on Neural Networks
1994–1998	Associate Editor, Signal Processing Magazine

Research Areas

Computational Learning Theory/Machine Learning/Pattern Recognition/Neural Networks
Numerical Optimization

Selected Recent Publications

M. Fugate, R. Christensen, D. Hush, and C. Scovel, "An equivalence relation between parallel calibration and principal component regression," submitted to *Journal of Chemometrics*, 2000.

D. Hush and C. Scovel, "Conditional performance bounds for machine learning," submitted to *Machine Learning*, 1999.

D. Hush and C. Scovel, "A new proof of concentration of Rademacher statistics," submitted to *Annals of Probability*, 1999.

D. Hush and C. Scovel, "On the VC dimension of bounded margin classifiers," to appear in Machine Learning, 2000.

D. Hush, "Training a sigmoid node is hard," *Neural Computation*, Vol. 11, pp. 1249–1260, 1999.

D. Hush and B. Horne, "Efficient algorithms for function approximation with piecewise linear sigmoidal networks," *IEEE Trans. Neural Networks*, Vol. 9, No. 6, pp. 1129–1141, 1998.